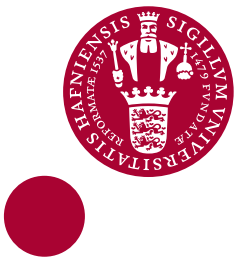

Genome and transcriptome of the porcine whipworm Trichuris suis

Jex, Aaron R.; Nejsum, Peter; Schwarz, Erich M.; Hu, Li; Young, Neil D.; Hall, Ross S.; Korhonen, Pasi K.; Liao, Shengguang; Thamsborg, Stig Milan; Xia, Jinquan; Xu, Pengwei; Wang, Shaowei; Scheerlinck, Jean-Pierre Y.; Hofmann, Andreas; Sternberg, Paul W.; Wang, Jun; Gasser, Robin B.

Published in:
Nature Genetics

DOI:
10.1038/ng.3012

Publication date:
2014

Document version
Publisher's PDF, also known as Version of record

nature genetics
OPEN

Genome and transcriptome of the porcine whipworm *Trichuris suis*

Aaron R Jex[1], Peter Nejsum[2], Erich M Schwarz[1,3], Li Hu[4], Neil D Young[1], Ross S Hall[1], Pasi K Korhonen[1], Shengguang Liao[4], Stig Thamsborg[2], Jinquan Xia[4], Pengwei Xu[4], Shaowei Wang[4], Jean-Pierre Y Scheerlinck[1], Andreas Hofmann[5], Paul W Sternberg[6,7], Jun Wang[4,8–11] & Robin B Gasser[1]

Trichuris **(whipworm) infects 1 billion people worldwide and causes a disease (trichuriasis) that results in major socioeconomic losses in both humans and pigs. Trichuriasis relates to an inflammation of the large intestine manifested in bloody diarrhea, and chronic disease can cause malnourishment and stunting in children. Paradoxically, *Trichuris* of pigs has shown substantial promise as a treatment for human autoimmune disorders, including inflammatory bowel disease (IBD) and multiple sclerosis. Here we report whole-genome sequencing at ~140-fold coverage of adult male and female *T. suis* and ~80-Mb draft assemblies. We explore stage-, sex- and tissue-specific transcription of mRNAs and small noncoding RNAs.**

Soil-transmitted helminths, including whipworm (*Trichuris*), hookworms (*Necator* and *Ancylostoma*) and the large roundworm (*Ascaris*), are among the most prevalent and devastating parasites of humans globally and predominate in impoverished nations[1]. *Trichuris* infects 1 billion people, and chronic infection of high intensity can lead to typhlitis, colitis, chronic dysentery and malnutrition through malabsorption as well as reduced physical and cognitive development[2]. Consequently, trichuriasis, which disproportionately affects children, has an estimated global burden of 1 million–6.5 million disability-adjusted life years, exceeding that of schistosomiasis, trachomiasis, trypanosomiasis or leishmaniasis[1]. Despite this, *Trichuris* species are classified by the World Health Organization as neglected parasites in urgent need of improved control[3].

Contrasting with the substantial burden of trichuriasis and other neglected helminths is the observation that human populations in endemic countries tend to suffer from substantially fewer immunopathological diseases[4], which are common and increasingly prevalent[5] in countries in which exposure to pathogens is limited. These observations have inspired the 'hygiene hypothesis'[6], which proposes that a lack of exposure of humans to common pathogens impairs immune function and leads to increased autoimmune disease. This hypothesis is supported by clinical data, with routine deworming positively[7] and early-childhood helminth infection negatively[8] correlating with autoimmune disorders. Recent studies have shown that porcine *Trichuris* (*T. suis*) administered to humans suffering from IBD (including Crohn's disease and ulcerative colitis) can reduce clinical symptoms[9,10]. Similar observations have been made in patients with multiple sclerosis[11]. Although helminths can alter immune responses in their hosts via a variety of excretory-secretory (ES) molecules[12], the specific interactions between *T. suis* and its host remain unclear. By sequencing the *T. suis* genome and transcriptomes (mRNAs and small RNAs), we provide deep insights into the molecular biology of this parasite and its modulation of host immune responses. These data provide a solid basis for exploring human trichuriasis, developing new anti-parasitic drugs and elucidating how helminths suppress autoimmune disorders.

RESULTS

Sequencing, assembly and synteny

We sequenced the genomes of single adult female and male *T. suis* at ~140-fold coverage, producing draft assemblies of 76 and 81 Mb, respectively (**Table 1**, **Supplementary Figs. 1** and **2** and **Supplementary Tables 1** and **2**). Matches to conserved eukaryotic genes indicated that each assembly is 96% complete, with minimal redundancy (**Supplementary Table 3**). Alignment of these assemblies showed high similarity, with 68 Mb aligning as direct one-to-one matches in blocks of a mean length of 2.5 kb. Overall, sequence identity was 99.2% (38,854 SNPs). Despite the reported XX and XY karyotypes for female and male *Trichuris*, respectively[13], we found no evidence for a Y chromosome among the male-specific scaffolds, suggesting that this chromosome contains largely repetitive sequences common to both sexes; this finding is consistent with observations made of the *T. suis* karyotype, suggesting that the sex chromosomes were the smallest chromosomal pair and were morphologically very

[1]Faculty of Veterinary Science, The University of Melbourne, Parkville, Victoria, Australia. [2]University of Copenhagen, Copenhagen, Denmark. [3]Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York, USA. [4]BGI-Shenzhen, Shenzen, China. [5]Eskitis Institute for Cell and Molecular Therapies, Griffith University, Nathan, Queensland, Australia. [6]Howard Hughes Medical Institute, California Institute of Technology, Pasadena, California, USA. [7]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA. [8]Department of Biology, University of Copenhagen, Copenhagen, Denmark. [9]Princess Al Jawhara Albrahim Center of Excellence in the Research of Hereditary Disorders, King Abdulaziz University, Jeddah, Saudi Arabia. [10]Macau University of Science and Technology, Taipa, Macau, China. [11]Department of Medicine, University of Hong Kong, Hong Kong. Correspondence should be addressed to A.R.J. (ajex@unimelb.edu.au), J.W. (wangjun@genomics.org.cn) or R.B.G. (robinbg@unimelb.edu.au).

Received 15 December 2013; accepted 22 May 2014; published online 15 June 2014; doi:10.1038/ng.3012

Table 1 Features of the scaffolded assembly of the adult male and female *T. suis* genomes

	Male genome	Female genome
k-mer (17 nucleotides) estimated genome size (in Mb)	83.6	87.2
Total read data; estimated coverage	11.73 Gb; 140×	12.36 Gb; 142×
Total scaffolded assembly size (Mb); total scaffolds	81.3; 60,856	76.0; 42,663
Total scaffolds of >200 bp: length (Mb); no. of scaffolds	74.2; 4,293	71.0; 3,288
Largest scaffold (Mb)	1.59	1.44
N50 in kb (scaffolds >200 bp)[a]	500	440
N90 in kb (scaffolds >200 bp); total number for >N90[a]	81.0; 185	104; 168
% GC content, whole genome; scaffolds >200 bp)	43.9; 43.6	43.6; 43.5
% repetitive sequence (scaffolds >200 bp)	31.7	32.3
Proportion of the genome that is coding, exonic; including introns	26.2; 66.2	29.2; 73.4
Number of protein-encoding genes	14,781	14,470
Mean gene length (kb)	3.7	3.9
Mean number of exons per gene	5.4	5.7
Mean exon length (bp)	271	270
Mean number of introns per gene	4.4	4.7
Mean intron length (bp)	511	509
Number of transfer RNAs	991	1,021

[a]N50 and N90 denote that 50% and 90% of assembly, respectively, is represented by scaffolds of at least this size.

similar in both sexes[13]. Repetitive sequences comprise ~32% of the genome, including 8% DNA transposable elements, 2.9% long tandem repeats and 3.3% retrotransposons (**Supplementary Table 4**). Each genome encoded ~1,000 transfer RNA (tRNA) genes, with copy numbers reflecting codon usage in protein-encoding regions (**Supplementary Fig. 3** and **Supplementary Tables 5** and **6**).

Protein-encoding gene set

The female and male *T. suis* genomes encode at least 14,470 and 14,781 protein-encoding genes, respectively, representing ~70% of each genome, including introns and exons. We identified 14,356 and 14,315 female and male genes with an ortholog or homolog in the opposite sex, with 10,403 genes being defined as unambiguous one-to-one orthologs. Evidence for sex-specific genes was limited, with just one and 41 supported as female and male specific, respectively (see **Supplementary Note**). Of these sex-specific genes, only three male genes have a predicted function, having homology to *C. elegans frk-1* (encoding a receptor tyrosine-kinase), *gpc-1* (encoding a G protein-coupled receptor (GPCR)) and *his-66* (encoding a histone protein), respectively. The sex-specific genes show no clustering among scaffolds, providing little evidence for their association with the sex chromosomes. Most of the remaining differences in the genes of the two genders relate to a higher copy number of some genes in the male. From both assemblies, we defined a unified set of 14,820 genes for *T. suis* (**Table 1** and **Supplementary Table 7**), with 12,910 (87.1%) supported by high-throughput RNA sequencing (RNA-seq) data. The majority (59.8%) of these genes have homologs (BLASTp cut-off: 1×10^{-5}) in other nematodes, including 6,286 (42.4%), 6,340 (42.7%), 6,149 (41.5%) and 8,480 (57.2%) in *Ascaris suum*[14], *Brugia malayi*[15], *Caenorhabditis elegans*[16] and *Trichinella spiralis*[17], respectively (**Fig. 1**). Functions were assigned to 9,342 (63.0%) protein-encoding genes (**Supplementary Tables 8–13**). Focusing on key functional or druggable proteins[14], we predicted 653 peptidases and 288 peptidase inhibitors. Peptidase classes S1 (116) and S8 (42) are expanded in *T. suis* compared with those represented in other nematode genomes[14–17]. The *T. suis* genome also encodes 269 phosphatases and 232 kinases. We identified a large complement of receptors and transporters[18]; these molecules include 228 GPCRs, as well as 1,962 channel, pore and transporter proteins. Among the last group are 133 peroxisomal protein importers, more than in *A. suum* ($n = 74$) (ref. 14), which suggests a greater importance of fatty-acid digestion and metabolism in *T. suis*.

We predicted 618 canonical ES proteins in adult *T. suis* (**Supplementary Tables 14** and **15**), including 165 proteases, many of which might have a role in disrupting intestinal epithelial cells in the host[19,20] and in the formation of the syncytial tunnel around the *Trichuris* stichosome[21]. Notable among these molecules are 33 chymotrypsin-like serine proteases, which have key roles in helminths associated with host invasion[22], immunosuppression[23] and tissue destruction[24]. In addition to proteases, non–membrane-bound transporters comprise a major component of the secretome. These transporters include 41 pore-forming toxins (porins), 25 of which have homology to the *Trichuris trichiura* porin TT47, which induces ion-conducting pores in planar lipid bilayers and assists in the formation of the syncytial tunnel in the intestinal epithelium[25]. Helminth-mediated immunomodulation by ES products is well documented[12]. Among the predicted *T. suis* ES proteins, we found a variety of immunomodulators (**Supplementary Table 16**). On the basis of these findings and available literature for helminths[12,26,27], we propose a *Trichuris*-driven immunomodulation model (**Supplementary Fig. 4**), in which the parasite suppresses inflammation by secreting (i) serpins to inhibit neutrophil cathepsins and elastases; (ii) apyrases to prevent conversion of regulatory T cells to pro-inflammatory T cells; (iii) cystatins to promote anti-inflammatory (producing interleukin-4 (IL-4) and IL-10) T cells by disrupting antigen presentation by dendritic and B cells; (iv) calreticulins that bind to dendritic cells and stimulate IL-4 production and limit inflammation by binding free calcium ions; and (v) molecular mimics[12] of host galectins, mammalian macrophage inhibitory factor and tumor growth factor-β that stimulate apoptosis in activated T cells, promote alternative activation of macrophages and block the stimulation of (proinflammatory) Toll-like receptor pathways. This model is consistent with the pathophysiology described for *Trichuris* infection[26], and probably operates in tandem with immunosuppressive processes linked to glycans[27] and lipids (for example, sphingolipids; see **Supplementary Note**).

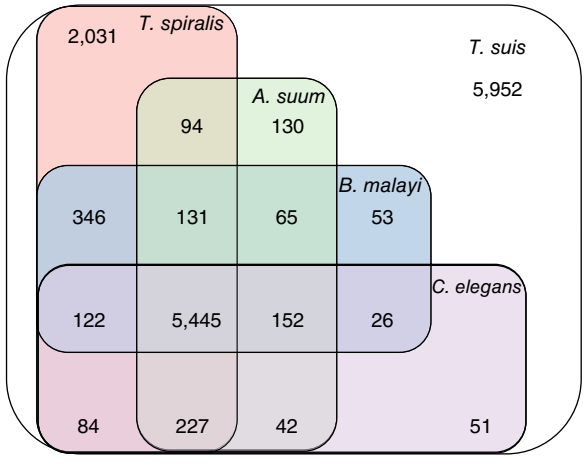

Figure 1 Homologs shared between *T. suis* (class Enoplea, order Trichocephalida) and related nematode species.

Transcriptome and differential transcription

We explored stage-, sex- and tissue-specific transcription (mRNAs), with a focus on parasite-host interactions. We predicted 36,763 transcripts, with 15,174 of them being perfect matches to the intron/exon chain (excluding UTRs), as annotated in the genome, and 21,589 representing novel splice isoforms (**Supplementary Fig. 5** and **Supplementary Table 17**). In total, 6,293 (43.7%) *T. suis* genes were predicted to encode at least two isoforms. The number of splice-isoforms per gene only moderately correlated with exon number (R^2 = 0.44; **Supplementary Figs. 6** and **7**). Alternative splicing correlated with gene function, reflected in an enrichment ($P \leq 0.05$; Pearson's chi-squared analysis) of protein catabolism (i.e., proteases), membrane-bound ion transport and kinase activity among alternatively spliced genes and apoptosis among single splice-isoform genes. Genes with multiple and single splice isoforms also differed in their conservation and predicted essentiality. Of the 6,307 *T. suis* genes with *C. elegans* homologs, alternatively spliced genes predominated by a three-to-one margin (4,510 versus 1,875) and represented 204 (75.8%) of the 269 essential genes predicted. The latter observation needs to be considered when mining helminth genomes for novel drug targets. If a novel inhibitor targeting products of such genes interacts with one of the spliced domains, isoform switching may be sufficient to overcome its effect.

Some protein domains were significantly associated ($P \leq 0.05$; Pearson's chi-squared analysis) with specific, alternative splice events, with exon skipping and the use of alternative first or last exons appearing to differ in their functional implications for transcription (**Fig. 2a**). Most notable was an over-representation of substrate-binding motifs (for example, immunoglobin, EGF-like or DnaJ) for genes biased toward transcripts with skipped exons. Remodeling of binding-motif structure through alternative splicing affects binding specificity in other organisms[28], and we propose that exon skipping is important in regulating binding specificity of proteins in *T. suis*. Given the varied functions associated with alternative first- or last-exon splicing events (**Fig. 2a**), we hypothesize that these specific modifications might play a part in regulating protein localization, another known role for alternative splicing[28].

T. suis undergoes substantial developmental changes throughout its direct life cycle[29]. To understand developmental processes in this parasite, we used RNA-seq to characterize transcription in various stages, sexes or body portions (stichosomal versus all non-stichosomal tissues from male and female adult worms): first and second (L1/L2), third (L3) and fourth (L4) larval stages adult male and female; stichosome and adult male posterior body and female posterior body excluding the stichosome (**Supplementary Fig. 8** and **Supplementary Table 18**). Overall, a number of major functional classes of proteins showed higher representation in the transcriptome of *T. suis* than in the genome (**Fig. 2b**). Secretory proteins were notable in this regard, making up ~4% of the *T. suis* gene set but representing ~10% of the transcriptional abundance in all libraries. Peptidases, particularly secreted peptidases, were also over-represented in the transcriptome and, notably, were upregulated during larval development and in the stichosome.

The stichosome is the thin, elongate anterior end of *Trichuris* embedded tightly within a syncytial tunnel[21] in the superficial layer of the large intestinal mucosa. Within this tunnel, the parasite secretes proteins and other molecules and absorbs nutrients from cell cytoplasm and surrounding tissue fluids, probably through thousands of bacillary cells[30]. Given its central importance in feeding and interaction with the host, we focused on transcription in the stichosome relative to the rest of the worm body (**Supplementary Table 18**; see **Supplementary Note** for detailed comparisons among other stages or tissues). Transcription was enriched for 2,210 genes (encoding 3,721 transcripts) in the stichosome relative to both the male and female posterior bodies (**Supplementary Table 18**). Among these genes are 160 peptidases (encoding 256 transcripts) and 41 porins (85 transcripts), supporting their role in host-tissue degradation and syncytial tunnel formation[19,25] (**Supplementary Fig. 9**). Also notable is the enrichment of a large number of secreted and

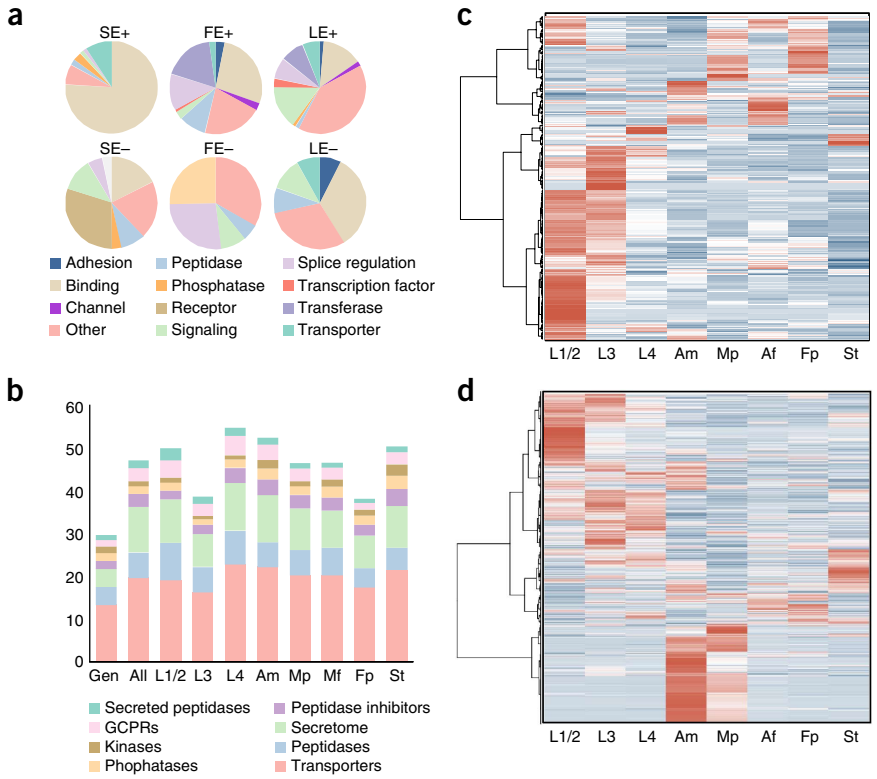

Figure 2 Stage- and tissue-specific small-RNA and mRNA transcriptome of *T. suis*. (**a**) Association between gene function and alternative splice variation. Charts show the inferred function of protein domains encoded by genes showing a statistically significant ($P \leq 0.05$; Pearson's chi-squared analysis) positive (+) or negative (–) bias toward skipped exon (SE), alternative first (FE) or last exon (LE) splice events. Only genes encoding ten or more transcripts are included in this analysis. (**b**) Proportional representation of major protein classes or groups encoded by the genome (Gen), and their proportional abundance in all transcriptomic data (All) and in larval (L1/2, L3 and L4), adult male (Am), female (Af) and tissue-specific libraries, including in the male (Mp) and female posterior body (Fp) and the stichosome (St). (**c**) Self-organizing heatmap (transcripts per million (TPM) values normalized by gene) clustering miRNAs by their transcription abundance (represented as $\log_2$-transformed reads per kilobase per million reads (RPKM) values) in each larval, adult and tissue-specific library. (**d**) Self-organizing heatmap (TPM values normalized by gene) clustering 22A-RNAs by their transcription abundance (represented as $\log_2$-transformed RPKM values) in each larval, adult and tissue-specific library.

membrane-bound transporters (222 genes encoding 371 transcripts) of various ions (for example, sodium, phosphate and calcium) and small molecules (for example, glucose and nucleosides). Sugar metabolism is enriched in the stichosome, suggesting that absorbed glucose is rapidly metabolized in the stichocytes. Also upregulated in the stichosome are transcripts associated with endocytosis and vesicle formation, lysozyme and peroxisome pathways as well as fatty acid and amino acid (cysteine and methionine; lysine) degradation. At least one isoform of each putative immunomodulatory gene encoded by *T. suis* is transcribed in the stichosome, with 22 transcripts encoding galactins, serpins, venom allergen–like proteins, apyrase or calreticulin specifically enriched in the stichosome relative to both the male and female posterior bodies.

Chymotrypsin-like (S1) serine proteases ($n = 28$ of 31 genes, and 51 of 135 transcripts) are also upregulated in the stichosome. Many are homologs of vertebrate plasmin, which is thought to regulate blood clotting in the host[31]. A poorly understood consequence of trichuriasis is bloody diarrhea[32], and some evidence suggests that *Trichuris* ingests blood[33]. It may be that some *T. suis* chymotrypsin-like serine proteases act as anticoagulants or assist in digesting blood, serum and tissue components (for example, fibrinogen). Notably, *T. muris* infection alters the mucus barrier in the host's gut epithelium, leading to an increased susceptibility to nematode infections[34] through the degradation of mucin 2 (Muc2) polymers[35]. Muc2 depolymerization by *T. muris* is blocked by chymostatin and antipain[35], suggesting a probable role for chymotrypsin-like and other serine proteases. Several of the secreted serine proteases enriched in the stichosome are homologs of *Schistosoma mansoni* serine protease 1 (SP1) and human kallikrein. The latter molecule regulates the degradation of kininogen to bradykinin, stimulating vasodilation, the cytosolic release of Ca^{2+}, neutrophil recruitment and increased inflammation[36]. SP1 is a potent vasodilator in mice[37], suggesting that it has an ability to convert vertebrate kininogen to bradykinin. Given the anti-inflammatory capacity of *T. suis*, we propose that some of these chymotrypsin-like serine proteases might degrade host kininogen but do not enable bradykinin production, thereby preventing bradykinin receptor stimulation and, thus, inhibiting inflammation. Notably, bradykinin receptors have key roles in various autoimmune disorders, including IBD[38] and multiple sclerosis[36].

Genetic regulatory networks

The *T. suis* gene set has complete RNA-interference machinery, suggesting potential for functional genomic studies and indicating a role for small noncoding RNAs in gene regulation. We explored these small RNAs in *T. suis* (**Supplementary Figs. 10** and **11** and **Supplementary Tables 19–21**) and produced ~435 million sequence reads. Approximately 92% of these reads mapped to the *T. suis* genome, with 16%, 23% and 9% classified as microRNAs (miRNAs), small interfering RNAs (siRNAs) and tiny noncoding RNAs (tncRNAs), respectively. Approximately 4% of the small-RNA reads mapped with an antisense (>80% of reads) bias to transposable elements, consistent with Piwi-interacting RNAs (piRNAs)[39]. However, similarly to small RNAs in *Ascaris suum*[40], < 0.01% had characteristics consistent with 21U-RNAs, which function as piRNAs in *C. elegans*[41].

We identified 319 miRNAs, with 132 having close homologs in other nematodes (**Supplementary Table 22**). These miRNAs accounted for 16% of all small-RNA reads sequenced, with tsu-let-7 (50% of all miRNA reads), tsu-miR-1 (17%), tsu-novel-51 (8%; a homolog of tsp-novel-51 miRNA from *T. spiralis*) and tsu-miR-228 (4%) the most highly transcribed. Approximately two-thirds of the miRNAs were most abundant in larval stages, suggesting a central role in development, with a diminishing number of miRNAs enriched in adults (**Fig. 2c**). This trend was reversed in the transition from L4 to adult female. To explore the functional implications of differential transcription of these miRNAs, we predicted miRNA-binding sites linked to 3′ UTRs among 22,954 of the 23,824 mRNA isoforms (representing 7,180 genes) for which at least part of the 3′ UTR could be identified on the basis of RNA-seq data. We focused on miRNA-mRNA interactions recognized or proposed for *C. elegans*. Of the 785,143 predicted binding sites with homology to *C. elegans* (both miRNA and mRNA), 300,042 were supported by information in public databases and 3,238 by experimental findings[42]. Owing to differences in gene copy number, these 300,042 binding sites represented 45 and 62 miRNAs as well as 3,205 and 3,877 coding genes in *C. elegans* and *T. suis*, respectively.

For *T. suis*, the shift from L3 to L4 coincides with a universal downregulation of 24 of these conserved miRNAs, including tsu-miR-1, tsu-miR-252 and tsu-miR-236—the second, seventh and eighth most abundant miRNAs, respectively, in *T. suis* overall. We identified 69 transcripts enriched in L4 (relating to 62 *T. suis* genes and 61 *C. elegans* homologs) with binding sites for each of these miRNAs. Many of the *C. elegans* genes with inferred homology to these transcripts are involved in larval or embryonic development (for example, *rol-3*, *slt-1* and *sox-3*), growth (for example, *egl-4*, *unc-44* and *lin-39*) or early sexual determination (for example, *sex-1*; WormBase), suggesting similar functional roles in *T. suis*. The maturation of *T. suis* to adulthood coincides with a variety of sex-specific changes in miRNA levels (relative to L4s). In both sexes, tsu-miR-228 (the fourth most abundantly transcribed miRNA in *T. suis*) and several isoforms of tsu-miR-61 were downregulated, and tsu-miR-34 and two 'minor' miRNAs—tsu-miR-256 and tsu-miR-50—were upregulated. Many of the coding genes ($n = 447$) upregulated in male and female adults are predicted to be co-regulated by tsu-miR-61 and tsu-miR-228. Homologs of these coding genes in *C. elegans* are enriched in GO terms (biological process) for embryonic and genital development, reproduction, morphogenesis and growth and metabolism, and include *tbx-2*, *vps-16*, *xnp-1* and *dyci-1* (WormBase). Notable among predicted tsu-miR-34–regulated genes were homologs of *srp-2* (encoding serpin-2, an anti-inflammatory protein in helminths)[12], which is downregulated in the adult worm (with the exception of the stichosome) compared with larval stages.

When we compared male and female *T. suis* adults, we found that major differences in miRNA transcription also related to tsu-miR-61, tsu-miR-228, tsu-miR-236 and tsu-miR-252, highlighting their importance in this nematode. Enriched in males are tsu-miR-228 and four copies of tsu-miR-61, and in females, tsu-miR-236, tsu-miR-252 and one copy of tsu-miR-61 (with closest homology to cel-miR-61-5p). Considering the ambiguity associated with the enrichment of different tsu-miR-61 isoforms in both males and females, we focused on tsu-miR-228, tsu-miR-236 and tsu-miR-252. In males, the enrichment of tsu-miR-228 coincides with a downregulation of 412 transcripts (representing 320 *T. suis* genes) with a predicted tsu-miR-228 binding site. On the basis of their function in *C. elegans* homologs, we infer many of these transcripts to be linked to vulva development (for example, *exc-4*, *mys-1*, *nekl-2* and *sem-4*), egg production (for example, *cbd-1*, *nsy-1*, *ppt-1*, *unc-29* and *unc-58*) and embryogenesis and germline development (for example, *bcat-1*, *lars-1*, *rnp-4*, *rpt-5*, *slt-1* and *tbp-1*). In females, enriched transcription of tsu-miR-236 and tsu-miR-252 coincides with a downregulation of 262 transcripts (representing 205 genes) predicted to be co-regulated by these miRNAs. Homologs of these 'female-suppressed' coding genes in *C. elegans* include genes involved in spermatogenesis (for example, *cogc-5* and *cpb-1*), male mating or fertility (for example, *goa-1* and *odc-1*), the regulation of germline specification or apoptosis (for example, *glp-1*, *him-1*, *rpt-5*, *let-60*, *vps-16* and *vps-41*) and chemosensation (for example, *crh-1*, *grk-2* and *lys-2*). Collectively, these data suggest that sexual dimorphism in *T. suis* might relate, at least

partially, to post-transcriptional sex suppression by miRNAs rather than exclusive transcriptional promotion by mRNAs.

In addition to miRNAs, we identified 1,028,808 putative small RNAs mapping to coding regions of the genome. Of these RNAs, 673,355 mapped antisense (≥80% of reads at each location) to exons, suggesting a potential role as siRNAs[41]. Most abundant among the siRNAs predicted for *T. suis* (except those derived from males) were sequences of 24–25 nt with a 5′ guanine (i.e., 24G and 25G), compared with 22G and 26G sequences predominating among siRNAs predicted for other nematodes to date[40,41]. Putative siRNAs were predicted for 3,497 protein-encoding genes. Many siRNAs have key roles in germline tissues[40,41]. In *T. suis*, we identified transcripts for 508 coding genes, for which putative siRNAs were uniquely transcribed in the adult female and the female posterior body relative to the stichosome, the adult male and the male posterior body (**Supplementary Tables 7** and **18**; see URLs). These coding genes were enriched for transposable elements/transposases, histones or histone methytransferases, DNA- or RNA-binding, chromatin folding and homeodomain-related proteins. Similarly, but in lower abundance, these functions were also enriched in relation to the 69 coding genes associated with siRNAs uniquely transcribed in the adult male and the male posterior body. We hypothesize that these highly transcribed siRNAs protect chromatin in the *T. suis* germline, and this hypothesis is supported by the observation that 162 of the female-enriched siRNAs are absent from the larval stages studied here (**Supplementary Tables 7** and **18**).

Novel class of tncRNAs

Conspicuous among the *T. suis* small RNAs is an abundance of 22-nt sequences with a 5′ adenine cap. Although representing just 2.9% (n = 58,307) of consensus small RNAs, these sequences represent 9.2% of all small-RNA transcription. By location, 22-nt 5′-adenylated sequences are evenly distributed between coding and noncoding regions, and within noncoding regions, between annotated (such as transposable elements, tRNAs or other noncoding RNAs) and un-annotated spaces. However, 89% of transcription attributed to these sequences relates to un-annotated, noncoding space in the genome. On the basis of their size and abundance, these sequences are consistent with tncRNAs[43]; however, they have characteristics not previously attributed to this class. For instance, in *T. suis*, they have a clear strand bias, with 83% of their transcription occurring on the Watson (i.e., antisense) strand. These 'antisense'-biased tncRNAs (henceforth called 22A-RNAs) form 1,208 clusters (ranging from 22 to 11,831 nt) among 238 assembly scaffolds, with a median of three 22A-RNA sequences per cluster at a median spacing of 97 bp. Although 40% of multicopy 22A-RNA sequences are found in the same cluster, clusters comprising one repeated 22A-RNA sequence are rare, and sequences are often shared among clusters and genomic scaffolds, indicating that tandem duplication is not the only mechanism associated with cluster formation. Few transposable elements are found within 25 kb of these clusters, suggesting that their insertion or translocation within the genome is not recent.

At this stage, we can only speculate about the function(s) and mechanism(s) of action of these sequences. Their 100-nt genomic neighborhoods vary: some regions resemble (but do not overlap with) known protein-encoding sequences, others resemble known noncoding RNAs such as tRNAs, and still others resemble neither. Eleven of these neighborhoods show partial similarities to cryptic tRNAs, ~100 nt in size, discovered in *C. elegans* by the modENCODE Consortium[44]. The sequences of the 22A-RNAs themselves are also heterogeneous, with only one over-represented 8-nt sequence motif (5′-A[CA]GATAT[GT]-3′) occurring in 4.5% (245 of 5,457) of 22A-RNA sequences (**Supplementary Fig. 12**). Given these findings, we propose that 22A-RNAs may be processed from larger noncoding RNAs of diverse types, some of which are highly conserved and familiar, others of which are both hypothetical and unfamiliar. Despite having no obvious promoter motif (such as that proposed for 21U-RNAs)[41], these sequences seem to be transcriptionally regulated, and their abundance varies substantially among stages, sexes and tissues in *T. suis*; including, notably, an enrichment in the adult male body and male posterior body relative to all other stages and tissues (**Fig. 2d**), which may suggest a role in the male germline. As a proportion of overall small RNA transcription, 22A-RNAs are most abundant in the stichosome, wherein they comprise ~22% of all small-RNA reads determined. Indeed, the stichosome is notably restricted in its classes of small RNAs, with miRNAs (39% of all small-RNA reads from the stichosome) and 22A-RNAs dominating the small-RNA population in this organ. Whether this finding points to an involvement of these novel *T. suis* noncoding RNAs in host interactions deserves detailed investigation.

DISCUSSION

Globally, helminthiases are seriously neglected causes of morbidity and mortality. Genomic and transcriptomic explorations of *T. suis* should enable the design of urgently needed therapeutics against human trichuriasis, one of the world's most important and neglected helminthiases. An intriguing feature of *T. suis* is its possible use as a therapy for human autoimmune disorders[9–11]. A detailed characterization of how this parasite modulates the host immune response is thus a key priority. Secreted proteins (including cystatins and serpins, thioredoxin peroxidase and various putative mimics of host proteins) seem to have a central role in this process, primarily through inhibiting inflammation. Our findings indicate a role for parasite-derived lipids, including the inferred synthesis of β-glucosylceramide, a known anti-inflammatory and putative therapy for IBD[45], during the early developmental phase of *T. suis* (see **Supplementary Note**). It is likely that both proteins and lipids work in concert with N-linked glycans, which are known immunomodulators produced by *T. suis*[27], particularly L4 and adult stages, in which pathways associated with their synthesis are transcriptionally enriched (see **Supplementary Note**). The detailed characterization of these molecules *in vitro* and *in vivo*, using existing models of IBD and other autoimmune disorders, might pave the way for parasite-derived therapies[5]. Indeed, a better understanding of the *T. suis*–host interactions might shed new light on why helminth exposure seems crucial for the development of a healthy immune system in humans. This is the first study to characterize the genomes of male and female individuals of a dioecious nematode. We found little evidence for sex-specific genes or assembly contigs, despite the reported XY karyotype of this species. However, intriguingly, miRNAs seem to have a major role in regulating sexual development in this species, with tsu-miR-228 in male, and tsu-miR-236 and tsu-miR-252 in female worms predicted to regulate and suppress key feminizing and masculinizing developmental genes, respectively. This is the first time that this has been observed for a metazoan.

URLs. WormBase, http://www.wormbase.org/; microRNA.org, http://www.microrna.org/microrna/home.do; RNA22, https://cm.jefferson.edu/rna22v1.0/; TargetScanWorm, http://www.targetscan.org/worm_52/; salient data files are accessible via http://gasser-research.vet.unimelb.edu.au/Trichuris_suis/ and ftp://ftp.wormbase.org/pub/wormbase/species/t_suis/; browsable male and female genomes are accessible via http://gasser-research.vet.unimelb.edu.au/jbrowse/JBrowse-1.11.2/index.html?data=TsuisMale/ and http://gasser-research.vet.unimelb.edu.au/jbrowse/JBrowse-1.11.2/

index.html?data=TsuisFemale/, respectively, or through WormBase via ftp://ftp.wormbase.org/pub/wormbase/species/t_suis/.

Accession numbers. All short-read data are available via Sequence Read Archive: SRR1041639, SRR1041640, SRR1041641, SRR1041642, SRR1041643, SRR1041644 (genomic DNA male); SRR1041645, SRR1041646, SRR1041647, SRR1041648, SRR1041649, SRR1041650 (genomic DNA female); SRR1041651, SRR1041652, SRR1041653, SRR1041654, SRR1041655, SRR1041656, SRR1041657, SRR1041658 (mRNA); SRR1041659, SRR1041660, SRR1041661, SRR1041662, SRR1041663, SRR1041664, SRR1041669, SRR1041670 (small RNA). Annotated assemblies of each genome are accessible via BioProject PRJNA208415 (male) and PRJNA208416 (female).

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We are indebted to the staff of the BGI-Shenzhen who assisted this study but whose names are not included in the authorship. We also acknowledge the continued contributions of staff at WormBase. This project was funded by the Australian Research Council (ARC), the National Health and Medical Research Council (NHMRC) of Australia and BGI-Shenzhen. This research was supported by a Victorian Life Sciences Computation Initiative (VLSCI) grant number VR0007 on its Peak Computing Facility at the University of Melbourne, an initiative of the Victorian Government (A.R.J. and R.B.G.). Other support to R.B.G. from the Alexander von Humboldt Foundation, Australian Academy of Science, the Australian-American Fulbright Commission, Melbourne Water Corporation and the IBM Research Collaboratory for Life Sciences – Melbourne is gratefully acknowledged. P.N. was supported by Danish Agency for Science, Technology and Innovation. N.D.Y. holds an NHMRC Early Career Research Fellowship. E.M.S. was supported by startup funds from Cornell University. P.W.S. is an investigator with the Howard Hughes Medical Institute (HHMI) and acknowledges support from the US National Institutes of Health (NIH).

AUTHOR CONTRIBUTIONS

P.N. and S.T. provided *T. suis*, and N.D.Y. purified nucleic acids for sequencing. L.H. and S.W. coordinated sequencing. A.R.J., S.L., P.K.K., R.S.H., J.X. and P.X. undertook the assembly, annotation and analyses of genomic and transcriptomic data. A.R.J., R.S.H., A.H., P.K.K. and E.M.S. planned and performed additional, detailed bioinformatic analyses. A.R.J., P.W.S., J.-P.Y.S. and R.B.G. drafted and edited the manuscript, tables, figures and supplementary information. A.R.J., N.D.Y., J.W. and R.B.G. conceived and planned the project. A.R.J., N.D.Y. and R.B.G. supervised and coordinated the research.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Hotez, P.J., Fenwick, A., Savioli, L. & Molyneux, D.H. Rescuing the bottom billion through control of neglected tropical diseases. *Lancet* **373**, 1570–1575 (2009).
2. Stephenson, L.S., Holland, C.V. & Cooper, E.S. The public health significance of *Trichuris trichiura*. *Parasitology* **121** (suppl.), S73–S95 (2000).
3. Anonymous. Schistosomiasis and soil-transmitted helminth infections (World Health Assembly Resolution WHA54.19) http://apps.who.int/gb/archive/pdf_files/WHA54/ea54r19.pdf (2001).
4. Okada, H., Kuhn, C., Feillet, H. & Bach, J.F. The 'hygiene hypothesis' for autoimmune and allergic diseases: an update. *Clin. Exp. Immunol.* **160**, 1–9 (2010).
5. Bach, J.F. The effect of infections on susceptibility to autoimmune and allergic diseases. *N. Engl. J. Med.* **347**, 911–920 (2002).
6. Elliott, D.E., Summers, R.W. & Weinstock, J.V. Helminths as governors of immune-mediated inflammation. *Int. J. Parasitol.* **37**, 457–464 (2007).
7. Endara, P. *et al.* Long-term periodic anthelmintic treatments are associated with increased allergen skin reactivity. *Clin. Exp. Allergy* **40**, 1669–1677 (2010).
8. Rodrigues, L.C. *et al.* Early infection with *Trichuris trichiura* and allergen skin test reactivity in later childhood. *Clin. Exp. Allergy* **38**, 1769–1777 (2008).
9. Summers, R.W., Elliott, D.E., Urban, J.F. Jr., Thompson, R.A. & Weinstock, J.V. *Trichuris suis* therapy for active ulcerative colitis: a randomized controlled trial. *Gastroenterology* **128**, 825–832 (2005).
10. Summers, R.W. *et al. Trichuris suis* seems to be safe and possibly effective in the treatment of inflammatory bowel disease. *Am. J. Gastroenterol.* **98**, 2034–2041 (2003).
11. Fleming, J.O. *et al.* Probiotic helminth administration in relapsing-remitting multiple sclerosis: a phase 1 study. *Mult. Scler.* **17**, 743–754 (2011).
12. Hewitson, J.P., Grainger, J.R. & Maizels, R.M. Helminth immunoregulation: the role of parasite secreted proteins in modulating host immunity. *Mol. Biochem. Parasitol.* **167**, 1–11 (2009).
13. Spakulová, M., Kralova, I. & Cutillas, C. Studies on the karyotype and gametogenesis in *Trichuris muris*. *J. Helminthol.* **68**, 67–72 (1994).
14. Jex, A.R. *et al. Ascaris suum* draft genome. *Nature* **479**, 529–533 (2011).
15. Ghedin, E. *et al.* Draft genome of the filarial nematode parasite *Brugia malayi*. *Science* **317**, 1756–1760 (2007).
16. *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
17. Mitreva, M. *et al.* The draft genome of the parasitic nematode *Trichinella spiralis*. *Nat. Genet.* **43**, 228–235 (2011).
18. Kaminsky, R. *et al.* A new class of anthelmintics effective against drug-resistant nematodes. *Nature* **452**, 176–180 (2008).
19. Cooper, E.S., Whyte-Alleng, C.A., Finzi-Smith, J.S. & MacDonald, T.T. Intestinal nematode infections in children: the pathophysiological price paid. *Parasitology* **104** (suppl.), S91–S103 (1992).
20. Drake, L.J., Bianco, A.E., Bundy, D.A. & Ashall, F. Characterization of peptidases of adult *Trichuris muris*. *Parasitology* **109**, 623–630 (1994).
21. Tilney, L.G., Connelly, P.S., Guild, G.M., Vranich, K.A. & Artis, D. Adaptation of a nematode parasite to living within the mammalian epithelium. *J. Exp. Zool. A Comp. Exp. Biol.* **303**, 927–945 (2005).
22. Toubarro, D. *et al.* Serine protease-mediated host invasion by the parasitic nematode *Steinernema carpocapsae*. *J. Biol. Chem.* **285**, 30666–30675 (2010).
23. Balasubramanian, N., Toubarro, D. & Simoes, N. Biochemical study and *in vitro* insect immune suppression by a trypsin-like secreted protease from the nematode *Steinernema carpocapsae*. *Parasite Immunol.* **32**, 165–175 (2010).
24. Toubarro, D. *et al.* An apoptosis-inducing serine protease secreted by the entomopathogenic nematode *Steinernema carpocapsae*. *Int. J. Parasitol.* **39**, 1319–1330 (2009).
25. Drake, L. *et al.* The major secreted product of the whipworm, *Trichuris*, is a pore-forming protein. *Proc. Biol. Sci.* **257**, 255–261 (1994).
26. Kringel, H., Iburg, T., Dawson, H., Aasted, B. & Roepstorff, A. A time course study of immunological responses in *Trichuris suis*–infected pigs demonstrates induction of a local type 2 response associated with worm burden. *Int. J. Parasitol.* **36**, 915–924 (2006).
27. Klaver, E.J. *et al. Trichuris suis*–induced modulation of human dendritic cell function is glycan-mediated. *Int. J. Parasitol.* **43**, 191–200 (2013).
28. Stamm, S. *et al.* Function of alternative splicing. *Gene* **344**, 1–20 (2005).
29. Beer, R.J. Studies on the biology of the life-cycle of *Trichuris suis* Schrank, 1788. *Parasitology* **67**, 253–262 (1973).
30. Sheffield, H.G. Electron microscopy of the bacillary band and stichosome of *Trichuris muris* and *T. vulpis*. *J. Parasitol.* **49**, 998–1009 (1963).
31. Hoover-Plow, J. Does plasmin have anticoagulant activity? *Vasc. Health Risk Manag.* **6**, 199–205 (2010).
32. MacDonald, T.T. *et al.* Histopathology and immunohistochemistry of the caecum in children with the *Trichuris* dysentery syndrome. *J. Clin. Pathol.* **44**, 194–199 (1991).
33. Burrows, R.B. & Lillis, W.G. The whipworm as a blood sucker. *J. Parasitol.* **50**, 675–680 (1964).
34. Hasnain, S.Z. *et al. Mucin* gene deficiency in mice impairs host resistance to an enteric parasitic infection. *Gastroenterology* **138**, 1763–1771 (2010).
35. Hasnain, S.Z., McGuckin, M.A., Grencis, R.K. & Thornton, D.J. Serine protease(s) secreted by the nematode *Trichuris muris* degrade the mucus barrier. *PLoS Negl. Trop. Dis.* **6**, e1856 (2012).
36. Golias, C., Charalabopoulos, A., Stagikas, D., Charalabopoulos, K. & Batistatou, A. The kinin system–bradykinin: biological effects and clinical implications. *Hippokratia* **11**, 124–128 (2007).
37. Da'dara, A. & Skelly, P.J. Manipulation of vascular function by blood flukes? *Blood Rev.* **25**, 175–179 (2011).
38. Marceau, F. & Regoli, D. Therapeutic options in inflammatory bowel disease: experimental evidence of a beneficial effect of kinin B1 receptor blockade. *Br. J. Pharmacol.* **154**, 1163–1165 (2008).
39. Ghildiyal, M. & Zamore, P.D. Small silencing RNAs: an expanding universe. *Nat. Rev. Genet.* **10**, 94–108 (2009).
40. Wang, J. *et al.* Deep small RNA sequencing from the nematode *Ascaris* reveals conservation, functional diversification, and novel developmental profiles. *Genome Res.* **21**, 1462–1477 (2011).
41. Ruby, J.G. *et al.* Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* **127**, 1193–1207 (2006).
42. Zisoulis, D.G. *et al.* Comprehensive discovery of endogenous Argonaute binding sites in *Caenorhabditis elegans*. *Nat. Struct. Mol. Biol.* **17**, 173–179 (2010).
43. Ambros, V., Lee, R.C., Lavanway, A., Williams, P.T. & Jewell, D. MicroRNAs and other tiny endogenous RNAs in *C. elegans*. *Curr. Biol.* **13**, 807–818 (2003).
44. Lu, Z.J. *et al.* Prediction and characterization of noncoding RNAs in *C. elegans* by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome Res.* **21**, 276–285 (2011).
45. Zigmond, E. *et al.* β-glucosylceramide: a novel method for enhancement of natural killer T lymphoycte plasticity in murine models of immune-mediated disorders. *Gut* **56**, 82–89 (2007).

ONLINE METHODS

Sample preparation and storage. *Trichuris suis* were isolated from experimentally infected pigs inoculated orally with a single dose of 5,000–50,000 embryonated eggs (Animal Ethics Permission No. 2010/561-1914; University of Copenhagen). Individuals of *T. suis* were isolated at 10 (L1/L2 larvae), 18 (L3s), 28 (L4s) and 49 (adulthood) d after inoculation (p.i.)[29,46] and washed in physiological saline (37 °C) and RPMI 1640 (GIBCO) with antibiotic-antimycotic (GIBCO). Adult male and female *T. suis* were separated. Stichosomes were excised from whole adult worms ($n = 10$, irrespective of sex), pooled and frozen, as were the posterior portions of the worms ($n = 10$ of each sex). All stages or tissues were snap frozen in liquid nitrogen and stored at −80 °C.

Genomic sequencing and assembly. Total genomic DNA was isolated each from a single adult male or female *T. suis*[47,48]. Paired-end (insert sizes, 170 bp and 500 bp) and mate-paired (800 bp, 2 kb, 5 kb and 10 kb) libraries were constructed from total and whole-genomic amplified (WGA) genomic DNA, respectively[14,49] and sequenced using a HiSeq 2000 machine (Illumina). Low-quality sequences, base-calling duplicates and adapters were removed using standard approaches. Sequence quality and heterozygosity were assessed by 17-mer frequency distribution[50] and genome sizes estimated[51]. Corrected and filtered data were assembled into contigs using SOAPdenovo v2.0 (ref. 50) and assessed for accuracy using SOAP2aligner[52]. Assembly completeness and redundancy were assessed using CEGMA[53] and RNA-seq data using Bowtie2 (ref. 54).

RNA isolation and RNA-seq. Total RNAs from L1/L2 ($n = 50{,}000$ from five pigs), L3 ($n = 15{,}000$ from four pigs) and L4 ($n = 3{,}000$ from two pigs), adult male ($n = 10$), adult female ($n = 10$), and stichosomal (mixed sex; $n = 10$) and nonstichosomal portions of adult females ($n = 10$) and males ($n = 10$) were individually purified using TriPure reagent (Roche). Polyadenylated (polyA+) RNA was purified from 10 µg of total RNA for each library using Sera-mag oligo(dT) beads and fragmented, purified and sequenced using HiSeq 2000 (refs. 14,49). Small noncoding RNAs (~18–30 nt) were isolated from 10 µg of total RNA for each library by size fractionation on polyacrylamide gels, purified, adaptor-ligated, reverse transcribed, amplified by PCR and sequenced using HiSeq 2000. All RNA-seq data were adaptor trimmed and length and quality filtered using standard approaches.

Synteny and polymorphism analysis, and annotation of repeat content. For comparative analysis, the assemblies for adult male or female *T. suis* were aligned using MUMmer3 (ref. 55). Repetitive sequences in each assembly were identified[14] using Tandem Repeats Finder (TRF)[56], RepeatMasker[57], LTR_FINDER[58], PILER[59] and RepeatScout[60], with a consensus population of predicted repetitive elements constructed in RepeatScout using fit-preferred alignment scores. Transfer RNAs were predicted using tRNA-SCAN[61]. The male assembly was explored for scaffolds likely to represent the male-specific Y chromosome[13,62]. Reads from all genomic sequence libraries each for male and female *T. suis* were aligned to their own and the opposite sex (both repeat unmasked and hard-masked) assembly (i.e., male-to-male, male-to-female, female-to-male and female-to-female) using Bowtie2 (ref. 54). Contigs with >80% coverage in same-sex but <20% coverage in opposite-sex read alignments were deemed 'sex-specific'.

Prediction and functional annotation of the protein-encoding gene set. The male and female protein-encoding gene set of *T. suis* was inferred in MAKER2 (ref. 63). Briefly, (i) the nonredundant *T. suis* transcriptome was aligned each assembly using BLAT[64] and filtered for full-length ORFs, which were used (ii) to train hidden Markov models (HMM) for *de novo* gene prediction using SNAP[65] and AUGUSTUS[66], with these models supplemented using (iii) homologous genes from *T. spiralis*[17] and *C. elegans*[16]; and (iv) all *T. suis* RNA-seq data from all libraries used to infer each transcript using Tophat2 (ref. 67) and Cufflinks2 (refs. 68,69); (v) all HMM-predicted, homology and evidence-based information was then combined into a single consensus gene set, and (vi) genes overlapping with predicted repetitive regions of the genome and/or having significant $E < 1 \times 10^{-5}$, BLASTn homology to known repetitive sequences (i.e., transposable elements) in RepBase[57] and no close homology to *C. elegans* or *T. spiralis* protein-encoding genes were removed. The male and female *T. suis* gene sets were unified by orthology prediction using InParanoid[70], with *T. spiralis*[17] as an out-group.

Conserved protein domains encoded by each gene were identified using InterProScan[71], with these data used to infer Gene Ontology[72]. Using Reciprocal BLASTp and OrthoMCL[73], the *T. suis* inferred proteome was clustered with predicted homologs or orthologs for other nematodes, including *Ascaris suum*[14], *Brugia malayi*[15], *C. elegans*[16] and *T. spiralis*[17]. Each contig was assessed for a known functional ortholog in the Kyoto Encyclopedia of Genes and Genomes (KEGG) using the KEGG orthology bases annotation system (kobas)[74]. In addition, *T. suis* inferred proteins were compared by BLASTx/BLASTp with protein sequences available for *A. suum*, *B. malayi*, *C. elegans* and *T. spiralis*, and in the databases UniProt[75], SwissProt and TREMBL[76], as well as specialist databases for key protein groups represented in MEROPS[77], WormBase[78], KS-SARfari and GPCR-SARfari, and the Transporter Classification database (TCDB)[79]. ES proteins were predicted using Phobius[80] and by BLASTp comparison with the validated signal peptide database (SPD)[81] and proteomic data for the nematodes *B. malayi*[82,83] and *Meloidogyne incognita*[84] and the trematode *Schistosoma mansoni*[85].

Differential transcription analysis of mRNA. Reconstruction and quantification (in fragments per kilobase per million reads (FPKMs)) of the *T. suis* transcriptome was conducted using TopHat2 (ref. 67) and Cufflinks2 (refs. 68,69). Predicted alternative splice events were classified[86]. Comparisons of splice events and gene function (based on encoded Pfam domains) were conducted by pairwise Pearson chi-squared analysis (P value ≤ 0.05). We also compared the relationship between gene essentiality and being a single or multi-isoform gene, with essentiality predicted[14]. Differential transcription was assessed using NOISeq[87], with 20% of the evaluated reads for each library used in five iterations to simulate technical replicates.

Annotation and differential transcription analysis of small noncoding RNAs. Canonical miRNAs were identified and quantified in miRDeep2 (refs. 88–90), and supported using miRNAs published for *A. suum*[40], *C. elegans*[91], *Haemonchus contortus*[92], *Brugia pahangi*[92] and *T. spiralis*[93]. The 3′ UTR for each Cufflinks-predicted *T. suis* transcript was identified by comparison with the *T. suis* genome annotation. Each 3′ UTR was screened for miRNA binding sites using PITA[94]. These binding sites were filtered on the basis of homologous miRNA-transcript binding interactions predicted for *C. elegans* in curated databases (microRNA.org, RNA22 and TargetScan) or demonstrated empirically[42]. All non-miRNA reads from each small-RNA library for *T. suis* were then aligned to the male *T. suis* genome using Bowtie2 (ref. 54) and clustered using ShortStack[95], with a minimum cluster depth cutoff of 10. Small-RNA reads having perfect alignment overlap (i.e., the same start and stop position) were defined as homologous and condensed into a consensus sequence by majority rule. Each consensus small RNA was classified[41] and nucleotide diversity within homologous small-RNA reads was assessed using custom Perl scripts. Specific small-RNA sequences (for example, 21U-RNAs)[41] and their 5′ and 3′ flanking regions were explored for sequence motifs using MEME[96]. Differential transcription among stage- or tissue-specific libraries was assessed (in reads per million mapped reads, RPKM) for miRNAs, siRNAs and 22A-RNAs using NOISeq[87], and clustered by stage- or tissue-specific transcription pattern using R.

Analysis of the genomic neighborhoods and primary sequences of 22A-RNAs. We extracted 22A-RNAs with 100-nt flanks lacking scaffolding (N) residues, merged those with spatial overlaps along the genome, and further condensed them to 80% sequence identity with CD-HIT-EST[97]. We probed resultant nonredundant 22A-RNA regions for protein-encoding exons with BlastX[79], and known noncoding RNAs (ncRNAs) with INFERNAL 1.1/cmscan[98]. BlastX was run against predicted proteomes from all published nematode genomes in WormBase WS240 (ref. 87), as well as the *T. suis* male proteome from this study; cmscan was run against the ncRNA database RFAM 11.0 (ref. 99). Regions passing these filters were tested for similarity to (i) genomic DNA from other nematode species, (ii) other 22A-RNA neighborhoods and (iii) novel ncRNAs from *C. elegans*. The first was assayed by BlastN against published nematode genomic sequences in WormBase WS240 (ref. 87). The second was assayed by BlastN against 22A-RNA regions spatially merged for

genomic overlaps but not condensed with CD-HIT-EST. The third was assayed by BlastN against a set of 8,126 *C. elegans* ncRNAs taken from the WS240 release of WormBase[87]. All searches used *E*-value thresholds of $\leq 10^{-3}$.

A set of 25,259 *C. elegans* ncRNAs was obtained from WormBase WS240. We filtered out those named 'asRNA', 'rRNA', 'scRNA', 'snoRNA', 'snRNA' or 'tRNA', leaving 8,126 ncRNAs with no official similarity to well-known structures. Most of these ncRNAs had been discovered by modENCODE[45]; 176 others were long noncoding RNAs[100]. We then checked for previously undescribed motifs via RFAM 11.0 and cmscan.

To discover whether 22A-RNA sequences contained novel motifs, we extracted 22A-RNA sequences without flanking protein-encoding or ncRNA similarities, merged them spatially and for 80% identity, and scanned with MEME[96], using a first-order Markov model from the adult male *T. suis* genome (via MEME's fasta-get-markov). We ran MEME with arguments: '-dna -revcomp -nsites 100 -bfile TS_M_200bpormore.1markov.txt -nmotifs 10 -evt 0.05 -minw 6 -maxw 22 -mod anr'. For one resulting 8-nt motif, we used FIMO[101] to determine where it occurred in original, unmerged 22A-RNA sequences, with arguments: '–bgfile TS_M_200bpormore.1markov.txt–output-pthresh 0.001'. The motif was displayed as a logarithmic WebLogo[102].

46. Kringel, H., Roepstorff, A. & Murrell, K.D. A method for the recovery of immature *Trichuris suis* from pig intestine. *Acta Vet. Scand.* **43**, 185–189 (2002).
47. Gasser, R.B. Molecular tools—advances, opportunities and prospects. *Vet. Parasitol.* **136**, 69–89 (2006).
48. Sambrook, J. & Russell, D.W. *Molecular Cloning: A Laboratory Manual* 3rd edn. (Cold Spring Harbor Laboratory Press, 2001).
49. Young, N.D. *et al.* Whole-genome sequence of *Schistosoma haematobium*. *Nat. Genet.* **44**, 221–225 (2012).
50. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
51. Lander, E.S. & Waterman, M.S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231–239 (1988).
52. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009).
53. Parra, G., Bradnam, K., Ning, Z., Keane, T. & Korf, I. Assessing the gene space in draft genomes. *Nucleic Acids Res.* **37**, 289–297 (2009).
54. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
55. Delcher, A.L., Salzberg, S.L. & Phillippy, A.M. Using MUMmer to identify similar regions in large sequence sets. *Curr. Protoc. Bioinformatics* **10**, 10.3 (2003).
56. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
57. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **25**, 4.10 (2009).
58. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
59. Edgar, R.C. & Myers, E.W. PILER: identification and classification of genomic repeats. *Bioinformatics* **21** (suppl. 1), i152–i158 (2005).
60. Price, A.L., Jones, N.C. & Pevzner, P.A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21** (suppl. 1), i351–i358 (2005).
61. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
62. Carvalho, A.B. & Clark, A.G. Efficient identification of Y chromosome sequences in the human and *Drosophila* genomes. *Genome Res.* **23**, 1894–1907 (2013).
63. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
64. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
65. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
66. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–W467 (2005).
67. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
68. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
69. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
70. O'Brien, K.P., Remm, M. & Sonnhammer, E.L. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* **33**, D476–D480 (2005).
71. Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**, W116–W120 (2005).
72. Harris, M.A. *et al.* The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
73. Fischer, S. *et al.* Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Curr. Protoc. Bioinformatics* **35**, 6.12 (2011).
74. Wu, J., Mao, X., Cai, T., Luo, J. & Wei, L. KOBAS server: a web-based platform for automated annotation and pathway identification. *Nucleic Acids Res.* **34**, W720–W724 (2006).
75. Wu, C.H. *et al.* The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* **34**, D187–D191 (2006).
76. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
77. Rawlings, N.D., Barrett, A.J. & Bateman, A. MEROPS: the peptidase database. *Nucleic Acids Res.* **38**, D227–D233 (2010).
78. Harris, T.W. *et al.* WormBase 2014: new views of curated biology. *Nucleic Acids Res.* **42**, D789–D793 (2014).
79. Saier, M.H. Jr., Yen, M.R., Noto, K., Tamang, D.G. & Elkan, C. The Transporter Classification Database: recent advances. *Nucleic Acids Res.* **37**, D274–D278 (2009).
80. Käll, L., Krogh, A. & Sonnhammer, E.L. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. *Nucleic Acids Res.* **35**, W429–W432 (2007).
81. Chen, Y. *et al.* SPD—a web-based secreted protein database. *Nucleic Acids Res.* **33**, D169–D173 (2005).
82. Bennuru, S. *et al. Brugia malayi* excreted/secreted proteins at the host/parasite interface: stage- and gender-specific proteomic profiling. *PLoS Negl. Trop. Dis.* **3**, e410 (2009).
83. Hewitson, J.P. *et al.* The secretome of the filarial parasite, *Brugia malayi*: proteomic profile of adult excretory-secretory products. *Mol. Biochem. Parasitol.* **160**, 8–21 (2008).
84. Bellafiore, S. *et al.* Direct identification of the *Meloidogyne incognita* secretome reveals proteins with host cell reprogramming potential. *PLoS Pathog.* **4**, e1000192 (2008).
85. Cass, C.L. *et al.* Proteomic analysis of *Schistosoma mansoni* egg secretions. *Mol. Biochem. Parasitol.* **155**, 84–93 (2007).
86. Wang, E.T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
87. Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. Differential expression in RNA-seq: a matter of depth. *Genome Res.* **21**, 2213–2223 (2011).
88. Friedländer, M.R. *et al.* Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.* **26**, 407–415 (2008).
89. Friedländer, M.R., Mackowiak, S.D., Li, N., Chen, W. & Rajewsky, N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.* **40**, 37–52 (2012).
90. Mackowiak, S.D. Identification of novel and known miRNAs in deep-sequencing data with miRDeep2. *Curr. Protoc. Bioinformatics* **36**, 12.10 (2011).
91. Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A. & Enright, A.J. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* **34**, D140–D144 (2006).
92. Winter, A.D. *et al.* Diversity in parasitic nematode genomes: the microRNAs of *Brugia pahangi* and *Haemonchus contortus* are largely novel. *BMC Genomics* **13**, 4 (2012).
93. Chen, M.X. *et al.* Identification and characterization of microRNAs in *Trichinella spiralis* by comparison with *Brugia malayi* and *Caenorhabditis elegans*. *Parasitol. Res.* **109**, 553–558 (2011).
94. Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U. & Segal, E. The role of site accessibility in microRNA target recognition. *Nat. Genet.* **39**, 1278–1284 (2007).
95. Axtell, M.J. ShortStack: comprehensive annotation and quantification of small RNA genes. *RNA* **19**, 740–751 (2013).
96. Bailey, T.L. & Elkan, C. The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **3**, 21–29 (1995).
97. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
98. Nawrocki, E.P. & Eddy, S.R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
99. Burge, S.W. *et al.* Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.* **41**, D226–D232 (2012).
100. Nam, J.W. & Bartel, D.P. Long noncoding RNAs in *C. elegans*. *Genome Res.* **22**, 2529–2540 (2012).
101. Grant, C.E., Bailey, T.L. & Noble, W.S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
102. Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).

doi:10.1038/ng.3012